动 物 学 研 究 2004, Feb. 25 (1): 73～80 CN 53-1040/Q ISSN 0254-5853
Zoological Research

DNA水平自然选择作用的检测

周 琦，王 文*
（中国科学院昆明动物研究所 细胞与分子进化重点实验室，中德马普青年科学家小组，云南 昆明 650223）

摘要： 上个世纪60年代，Kimura提出的“中性进化”假说使经典的达尔文自然选择学说遭遇了前所未有的挑战。但新近的研究表明：在DNA水平，越来越多的证据支持“自然选择”的进化理论。这些研究成果得益于近年来大量群体和基因组DNA数据的积累，以及理论群体遗传学的发展。在DNA水平检测选择作用是否存在的方法包括两大类：种内多态性检验和种间差异度检验。前者以Tajima（1989）提出的*D*检验为代表，后者大都基于“中性条件下，种内与种间进化速率一致”的原理。这些方法以中性假说作为零假设，结合统计检验方法分析DNA数据，被称为“中性检验”。这些方法对于解决一些有关进化的基础理论问题和人类遗传学及生物信息学的深入研究都具有重要意义。本文介绍几个应用广泛的检测方法，以使国内的读者了解它们的基本思路和操作方法。

关键词： 自然选择；DNA；统计检验
中图分类号： Q31；Q111 **文献标识码：** A **文章编号：** 0254-5853(2004)01-0073-08

Detecting Natural Selection at the DNA Level

ZHOU Qi, WANG Wen*
(*CAS-Max Planck Junior Scientist Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, the Chinese Academy of Sciences, Kunming* 650223, *China*)

Abstract: In the 1960s, the neutral theory proposed by Kimura caused an unprecedented challenge to the classical Darwin's theory of natural selection. However, recent advances in evolutionary genetics have provided a great deal of evidence on the role played by natural selection at the DNA level. These progresses have been stemmed from the appearance of enormous DNA sequence data of populations and genomes as well as the development of theoretical population genetics. There are mainly two kinds of approaches to detect selection at the DNA level: intraspecific polymorphism tests and interspecific divergence tests. The former one is represented by Tajima's (1989) *D* test while the latter one is based on the principle that the intraspecific polymorphism should be consistent with the interspecific divergence under neutrality. These methods are usually called "neutrality tests" because the neutrality hypothesis is taken as the null hypothesis in the tests. They are important tools not only in solving the basic theoretical questions in evolutionary biology but also in interpreting data and results obtained in the studies of human genetics and bioinformatics. In this paper, we shall review the progresses in detecting natural selection at the DNA sequence level and introduce the basis and application of several widely used tests.

Key words: Natural selection; DNA; Statistical test

进化论作为19世纪最重要的四项科学发现之一，其物竞天择的理论大大改变了人们以往对自然界的认识。但是20世纪60年代蛋白电泳技术(Hubby & Lewontin，1966)的发明和应用，使人们在氨基酸水平观测到了大量超出预期的遗传变异。当时这一结果与达尔文进化论的核心内容——自然选择在进化过程中起主导作用的观点所预期的现象相矛盾。之后Kimura在1968年提出了中性进化理论，并对这一矛盾做出解释。该理论认为，基因中的变异大多为中性，基因漂变、种群大小的变化和

收稿日期：2003-08-05；接受日期：2003-10-21
基金项目：中德马普青年科学家小组基金资助
*通讯作者 (Corresponding author)，E-mail：wwang@mail.kiz.ac.cn

种群迁徙等随机事件是决定进化的关键因素 (Kimura, 1968, 1983)。由于蛋白质的多态性和分子钟 (molecular clock) 现象都为中性假说提供了有力的证据，进化生物学一度分裂成分子水平的中性进化论和宏观角度的现代达尔文论的选择主义。

尽管如此，自然选择论依然影响深远。自 1983 年 DNA 测序技术用于群体遗传研究以来 (Kreitman, 1983)，在 DNA 水平上寻找自然选择作用成了进化生物学界的一个热点问题。一系列的重要研究逐步发现选择在 DNA 分子水平的重要作用。如 Fay et al (2002) 系统分析了果蝇等的基因组后，认为选择作用在形成现有基因变异格局中占有主导地位。在新基因的起源发生过程中，选择作用也承担了重要角色。如在 *jingwei* (精卫) 和作者发现的 *sphinx* (司芬克斯) 基因中，基因在自然选择的推动下发生了快速进化，从而很快形成了新的功能 (Long & Langley, 1993; Wang et al, 2000, 2002a)。我们也在人类与雄性生殖相关的基因中检验到了选择的存在 (Wyckoff et al, 2000)。上述研究成果为自然选择提供了实验证据。随着研究的逐步深入，在 DNA 水平上进行选择作用检验的理论和方法也已得到很大的完善。这些方法大都是以中性假说为零假设，通过统计学的检验方法分析 DNA 数据，进而检验自然选择的存在。它们被统称为中性检验 (neutrality test)。

这些检验方法的应用，不仅为进化生物学，同时也为人类遗传学和生物信息学提供了强有力的 DNA 数据分析工具和完善的逻辑支持。如一个抗性或疾病基因是否受自然选择作用而产生动态变化，基因中现行变异格局是否为选择所致等等这些生物学家所关心的问题，都可以运用这些方法进行分析推断。遗憾的是，国内对这一领域的认识和在这方面的工作尚处于起步阶段，因此阻碍了相关学科的发展，特别是人类遗传学和基因组学的发展。目前有必要对这些方法的原理和操作进行一些综合性的介绍。

本文将对六个检验选择作用的常用方法进行介绍，并指出其存在的问题和可能的发展。下文涉及的有关中性检验的计算都可以由相应软件①完成。依据用于分析的 DNA 数据类型，这些方法可分为利用种内多态性比较 (intraspecific polymorphism) 的检验和利用种间分歧度比较 (interspecific divergence) 的检验。在实验分析的过程中，单独采用某一种检验往往不能得到一个可靠的推断，经常需要同时运用几种检验，同时综合生物学背景，才能得到一个比较可信的结果。

1 基于种内多态性的中性检验

基因的长期进化和遗传多态性是同一个进化过程中的两个不同的层面 (Ohta & Kimura, 1971)。遗传多态性往往是研究进化动力学的切入点。与蛋白电泳相比，DNA 的多态性包含了更丰富和更本质的遗传进化信息，并且在操作上有大量连锁的所谓中性位点可供检视。所以，DNA 多态性已成为现代进化遗传学研究的首选材料。

衡量 DNA 多态性比较常用的参数有三种：分离位点数 (segregating sites)，以 K 表示，指所取 DNA 样本中具有不同碱基序列的位点数目；任意两序列之间核苷酸差异的平均数 (the mean number of nucleotide differences)，以 Π 表示；变异的频率谱线 (frequency spectrum)，意指变异在时间上先后出现的在不同谱系中的分布差异，或是根据变异出现的频率计算的杂合度。

Tajima (1989) 提出第一个基于种内 DNA 序列比较的中性检验方法后，一系列类似的检验方法应运而生。它们从原理和构建的方法上都沿袭了 Tajima's D 检验的思路：首先通过上文提到的几个衡量 DNA 多态性的参数对群体遗传参数 θ 进行估计。此处的 θ 是描述种群动态的参数，理论值为 4 $N_e\mu$ (N_e 为有效种群大小，μ 为突变速率)。在中性假说的条件下，不同方法产生的 θ 的估计值应该相等。基于此理念构建单侧检验或双侧检验统计检验式：

$$\frac{(L_1 - L_2)}{\sqrt{V(L_1 - L_2)}} \qquad (1)$$

式中 L_1、L_2 为通过不同方法产生的 θ 的估计值，具有相同的统计学期望值；V 代表方差计算。

通过蒙特卡罗随机模拟 (Monte-Carlo simulation) 产生关于统计检验的分布曲线和临界值，用于结果检验。(1) 式若得到显著 (significant) 的结果，则表示种群受到了基因随机漂变以外的因素影响，偏离了中性模型。

①MEGA2(http://www.megasoftware.net)、PAML(ftp://abacus.gene.ucl.ac.uk/pub/paml/)和 DNAsp(http://www.ub.es/dnasp/)。

1.1 Tajima's D 检验

这是Tajima（1989）提出的中性检验方法，通过分离位点数目 K 和任意两条序列之间差异的平均数 Π 对 θ 进行估计：

$$E(K) = a_1\theta_w \quad (\text{Watterson}, 1975) \qquad (2)$$

其中 $a_1 = \sum_{i=1}^{n-1}\frac{1}{i}$，$n$ 为所研究样本数目；

$$E(\Pi) = \pi \quad (\text{Tajima}, 1983) \qquad (3)$$

然后根据（2）和（3）构建Tajima's D 检验：

$$D = \frac{\pi - \theta_w}{\sqrt{V(\pi - \theta_w)}} \quad (\text{Tajima}, 1989) \qquad (4)$$

式中 $\theta_w = \frac{K}{a_1}$。

蒙特卡罗随机模拟显示，D 值的分布并非左右对称的正态分布，反而与 β 分布比较接近。所以，Tajima建议实际操作过程中，用 β 曲线来拟合 D 值曲线。

在用于检验的参数中，由于 K 的计算不考虑突变位点的频率，只计算位点的数目。所以即使是整个种群中所占比例很低的变异（mutations of low frequency）也将对 K 值产生很大影响。而 Π 由于计算的是序列差异的平均值，因此对平均杂合度贡献很小的低比例变异不会对 Π 产生很大的影响（图1）。

Seq1：AGCCCTTAC<u>G</u>
Seq2：AGCCCTTA<u>C</u>C
Seq3：AGC<u>T</u>CTTATC
Seq4：A<u>T</u>CTCTTATC

图1　K 和 θ 的计算
Fig.1　Computation of K and θ

假设有如图4条DNA序列，因为4个样本中有4个位点的碱基不同，所以 $K=4$；假设第 i 条序列与第 j 条序列间的差异是 d_{ij}，则：$d_{12}=1$，$d_{13}=3$，$d_{14}=4$，$d_{23}=2$，$d_{24}=3$，$d_{34}=1$，$\Pi=(d_{12}+d_{13}+\cdots+d_{34})/6=2.33$，$\theta_w=2.18$。

In the four DNA sequences listed above, there are 4 sites that have different nucleotide. That is to say, $K=4$. If the difference between the ith sequence and the jth sequence is dij, then: $d_{12}=1$, $d_{13}=3$, $d_{14}=4$, $d_{23}=2$, $d_{24}=3$, $d_{34}=1$, $\Pi=(d_{12}+d_{13}+\cdots+d_{34})/6=2.33$, $\theta_w=2.18$.

利用 K 和 Π 的特征差异，当所研究种群中出现有害变异（deleterious mutation）时，这些变异将由于受负选择作用（negative selection 或 purifying selection）而在种群中保持比较低的比例。种群中低比例的变异将比中性条件下有所增加，DNA数据上体现的效果为 θ_w 值增大，得到负的 D 值。当种群中的某一条等位基因受到强烈的正选择作用（positive selection）时，其附近与之紧密连锁的座位上的中性甚至轻微有害的变异，将伴随这条被选择影响的等位基因比例的升高而相应提升在种群中的比例，这样的现象被称为搭载效应（hitchhiking）或选择扫荡（selective sweep）。巧合的是，搭载效应过后，中性突变的积累同样将造成额外的低比例变异。因此，D 检验如果得到负的显著结果，既有可能是负选择造成的，也有可能是搭载效应的信号。而反之，当种群受到平衡选择（balancing selection）的作用时，群体中会存在两条或几条频率较高的等位基因，D 将为正值。例如最近我们在果蝇4号染色体中发现的情况（Wang et al，2002b）与平衡选择相符。果蝇4号染色体自摩尔根时代就被认为没有重组，因而在进化上可视为一个单元，成为当代分子进化遗传学中检验选择的一个经典案例。然而我们的研究表明：4号染色体实际上存在着重组。在一个约200 kb的区域内，二型性（dimorphism）明显地存在，其Tajima's D 值为正值。暗示着这一区域有可能存在平衡选择作用。

值得注意的是，选择并非造成 D 值显著的唯一原因，瓶颈效应和大规模的碱基插入或缺失也可能造成 D 小于零（Tajima，1989）；此外 D 值本身也并非呈现严格的 β 分布，用 β 曲线检验结果将会产生一些误差，因此必要时还需另作电脑模拟（Simonsen et al，1995）。

1.2 Fu和Li的检验方法

上个世纪80年代由Kingman（2000）、Tajima（1983）和Hudson（1983）等人奠定基础并迅速发展起来的种群遗传学溯祖理论（coalescent theory），使人们可以通过DNA样本对种群的进化历程进行动态追踪。Fu & Li（1993）提出的Fu和Li检验正是运用了这一理论，对变异在不同进化时间上的分布情况进行比较，进而检验种群进化是否符合中性模型。

在系统树中，与底层DNA样本序列直接相连的枝上的突变定义为外缘突变（external mutation；图2：d，e，f，g，h）；相对地，不直接与DNA样本序列相连的枝上的突变为内部突变（internal mutation；图2：a，b，c）。从系统树可见，外缘突变在时间上比内部突变晚或者说接近现代。

如果种群受到负选择作用，有害变异频率因选择而降低；或是某一条有益的等位基因在种群中的频率由于受正选择作用刚固定不久，都会导致外缘

图 2 一棵 5 个 DNA 序列构成的系统树

Fig.2 A phylogenic tree made from five DNA sequences

摘自 Fu & Li（1993）。图示每一个结点代表两条 DNA 序列的共同祖先，由上至下意味着时间上的由古至今。t_m（$m=2, \cdots, 5$）代表由 m 条序列回溯至（$m-1$）条序列所需代时（generation time）。

From Fu & Li (1993). Every node represents the common ancestor of two DNA sequences. t_m ($m=2, \cdots, 5$) is the time (number of generations) required for the coalescence from m sequences to m-1 sequences.

突变相对内部突变数量大大增多。反之若是受到平衡选择作用，则外缘突变会比较少。Fu 和 Li 通过溯源法对外缘突变和内支突变的期望值进行比较，若是在中性模型下，两者应该没有差异。基于此，构建了四个检验方法。这里仅就其中常言的根据含有外群的系统树构建的检验作出介绍。

外群（outgroup）是指与所研究物种或分类元近缘但在进化上又不属于所研究类群的分类元。构建一棵有外群的系统树，即有根树（rooted tree），在 Fu 和 Li 检验中用于估计外缘突变。经过 Fu 和 Li 的推导：

$$E(\eta_e) = \theta \quad \text{(Fu \& Li, 1993)} \tag{5}$$

$$E(\eta_i) = (a_1 - 1)\theta \quad \text{(Fu \& Li, 1993)} \tag{6}$$

其中 η_e 代表外缘突变数目，η_i 代表内支突变数目；$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$，$n$ 是所取样本数目。

根据（5）和（6）构建：

$$G = \frac{\eta_e - \frac{\eta_i}{a_1 - 1}}{\sqrt{V(\eta_e - \frac{\eta_i}{a_1 - 1})}} \quad \text{(Fu \& Li, 1993)} \tag{7}$$

类似地，G 的分布近似 β 分布。负的 G 值暗示种群有可能受负选择作用，反之则暗示有可能是平衡选择作用。Fu 和 Li 系列还有三个根据无外群系统树构建的检验，只是引用不同方法对 θ 进行估计，这里不一一赘述。

Fu（1997）同时指出，Fu 和 Li 检验在检验中性假说方面有可能比 Tajima's D 检验更灵敏。主要基于以下原因：第一，Fu 和 Li 检验融合了溯源法，对突变在时间上的先后有了动态的追踪，而 Tajima's D 检验仅考察某一时间点多态性的数据。第二，Fu 和 Li 检验对 θ 的估计用的是 η_e 和 η_i，Tajima's D 检验用的是 K 和 Π，通过计算两者的协方差值可知，K、Π 之间的关联程度大于 η_e、η_i（图 3）。从统计学角度来说，用于检验的变量之间关联越小，检验分辨率越高。

1.3 Fay 和 Wu 的 *H* 检验

如前所说，搭载效应是在没有或极少发生重组的区域，中性或轻微有害的变异因为与某一受正选择作用的变异连锁而在种群中比例升高的现象。在进行 DNA 数据分析时，进化生物学家面临的一个很大问题就是，往往不同的进化因素将产生相似或相同的 DNA 多态现象。例如，背景选择（background selection，由于某一基因受负选择作用，与

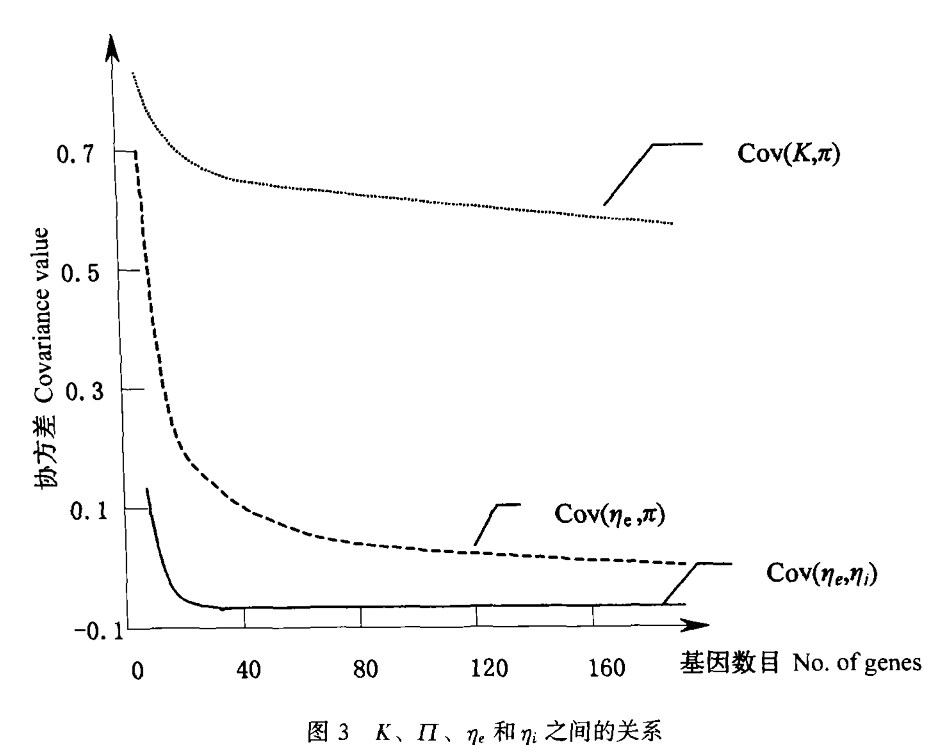


图 3　K、Π、η_e 和 η_i 之间的关系

Fig.3　Relationships among K, Π, η_e and η_i

摘自 Fu & Li（1993）。最高的点线为 K 和 Π 协方差线，中间的点划线为 η_e 和 π 协方差线，最低的实线为 η_e 和 η_i 的协方差曲线。
From Fu & Li (1993). The dotted curve represents covariance of K and Π; the dashdotted curve is the covariance of η_e and π; the solid curve is the covariance of η_e and η_i.

之连锁的变异比例降低的现象）与搭载效应都将造成种群平均杂合度的降低。此时一些中性检验将在区分两者之间显得力不从心。

为解决这一问题，Fay & Wu（2000）提出了一个专职检验搭载效应的中性检验方法：H 检验。H 检验与 Tajima's D 检验非常相似，所不同的是后者用了通过 K 估计得到的 θ（θ_w）与通过 Π 估计得到的 θ（θ_Π）进行比较；而 H 检验用通过变异频率估计得到的 θ（θ_H）与 θ_Π 比较。假设在 n 条染色体中，出现过 i 次的变异的数目为 S_i，则：

$$\theta_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \quad \text{(Fay \& Wu,2000)} \quad (8)$$

θ_H 对于高比例的变异（mutations of high frequency）比较敏感。当有搭载效应存在时，将产生异常高比例的变异，这是搭载效应区别于背景选择的一个显著标志。利用这一特征构建 H 检验：

$$H = \frac{\theta_H - \theta_\pi}{\sqrt{V_{ar}(\theta_H - \theta_\pi)}} \quad \text{(Fay \& Wu,2000)} \quad (9)$$

当 H 值在统计学上显示显著结果时，即暗示所研究种群有可能受搭载效应的影响。Fay 和 Wu 将这一方法用于分析果蝇编码附属腺体微管蛋白的基因 *Acp26Aa* 的 DNA 序列。在这一段约 350 bp 的 DNA 序列中，发现了比中性条件下异常多的高比例变异以及平均变异程度的降低。暗示这一区域内存在搭载效应。

2　基于种间差异数据的中性检验

中性假说预期，中性突变的随机遗传漂变是进化的主要推动力，所以种内 DNA 多态性和种间 DNA 分歧度的进化速率应该一致，即种内多态性与种间分歧度在中性模型下应当成比例。若通过检验，种内多态性与种间分歧度存在显著的偏差，则暗示选择作用的存在。

有必要强调的是，之前提到的检验方法都是对“严格中性假说”（strictly neutrality）的检验，即物种每一个变异都被认为是中性的。下面将要提到的几个基于种间数据的检验都是对 Kimura 中性假说（Kimura's neutral theory）的检验，即大部分的变异是中性的；所以后者的假设比前者弱很多。

2.1　McDonald 和 Kreitman 检验（MK 检验）

同义突变（synonymous substitution）指不改变氨基酸序列的突变。错义突变（replacement）指改变氨基酸序列的突变。MK检验的原理是：在无选择作用的中性条件下，所研究基因的种内的同义、错义突变应与种间同义、错义突变成正比。反之，则推翻零假设，即基因在不同物种中受到了选择的作用。MK检验思路简洁，计算简单，但在检验中性假说方面却很有说服力。而且该检验与以上提到的检验相比，不需要很多假设限制，重组和种群大小的动态对检验结果没有影响，所以应用广泛。

McDonald & Kreitman（1991）对所研究的DNA序列的位点首先进行分类，以区分种内差异和种间差异。将种内个体间无碱基差异而种间有明显碱基差异的位点，定义为固定位点（fixed site），作为种间差异的标志。将种内个体间有碱基差异的位点，定义为多态性位点（polymorphic site），作为种内多态性的标志。分辨出样本的多态位点和固定位点之后，将各位点上的突变再按同义突变位点和错义突变位点加以区分。按照MK检验的原理，在中性条件下：

$$\frac{E(n_f)}{E(s_f)}=\frac{E(n_p)}{E(s_p)} \quad \text{(McDonald \& Kreitman, 1991)} \qquad (10)$$

式中 n_f 代表既是错义突变位点又是固定位点的位点数，s_f 代表既是同义突变位点又是固定位点的位点数，n_p 代表既是错义突变位点又是多态位点的位点数，s_p 代表既是同义突变位点又是多态位点的位点数。

当选择作用存在于不同物种中时，（10）式两边会不相等。此时，可用统计学的 G-test 检验等式两边比例差异的显著性。若显著，也就是说物种间的错义突变数目大于基于种内多态性估计得到的期望值，说明基因在物种间受到了选择作用。

表1列出McDonald & Kreitman（1991）对果蝇的三个种，*D. melanogaster*、*D. simulans* 和 *D. yakuba* 的 *Adh* 基因序列的分析结果。就表1的数据为例，7/17与2/42显然差异显著（$G=7.43$，$P=0.006$）。暗示着该基因座位上选择作用的存在。

2.2 Hudson-Kreitman-Aquadé 的检验方法（HKA检验）

该检验方法基于的原理与MK检验相近，但运用的是统计学的卡方（χ^2）检验。即计算出种间和种内差异的卡平方和，再检验实验结果是否与中性条件下的期望值吻合，所以在统计学上也被称为吻合度检验（goodness of fit test）。

假设 K_{1i} 代表种1内第 i 座位DNA序列的分离位点数目，K_{2i} 代表种2内第 i 座位DNA序列的分离位点数目，D_i 代表种1和种2间第 i 座位序列的碱基差异数。将三者的卡平方和相加得到：

$$\chi^2=\sum\frac{[K_{1i}-E(K_{1i})]^2}{V(K_{1i})}+\sum\frac{[K_{2i}-E(K_{2i})]^2}{V(K_{2i})}+\sum\frac{[D_i-E(D_i)]^2}{V(D_i)} \quad \text{(Hudson et al, 1987)} \qquad (11)$$

Kreitman & Hudson（1991）将（11）用于果蝇 *Adh* 基因和5'侧翼序列（flanking region）两个区域DNA序列的比较检验。5'侧翼序列的每一个位点的突变都是同义突变，因此可假定为一段中性突变区域。*Adh* 基因与该区域比较得到显著的 χ^2 值（$P=0.05$），显示 *Adh* 基因序列上的变异不符合中性模型，暗示着 *Adh* 基因上存在着选择作用。

HKA检验对数据的要求比较高。计算 K 时需要有两个物种，并需要有两个或两个以上座位的DNA数据。其次该检验要求所研究种群大小保持恒定不变，座位间无连锁。

2.3 K_a vs. K_s 检验（Z 检验）

自然界中发生的很多错义突变都是有害突变。在这些突变位点上，碱基的替换将由于负选择作用而保持比较低的突变速率。假设 K_a 为错义突变速

表1 种内和种间的同义突变和错义突变数目

Table 1 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	固定位点（种间差异）Fixed site	多态位点（种内多态）Polymorphic site
错义突变 Replacement	7	2
同义突变 Synonymous	17	42

摘自 McDonald & Kreitman（1991）。
From McDonald & Kreitman（1991）.

率，K_s 为同义突变速率。由于同义突变不改变氨基酸序列，因此可假定同义突变为中性突变。大部分情况下，DNA 序列的 K_a/K_s 值由于负选择作用而小于 1。在中性条件下，K_a/K_s 值期望值为 1。但当正选择作用存在时，某一受正选择作用的等位基因的 K_a/K_s 将升高，甚至显著大于 1。这时可通过 Z 检验（单侧检验）来判断 K_a 和 K_s 之间是否存在显著差异，若 K_a 显著大于 K_s，即为正选择的标志。由于 Z 检验所要求的 DNA 序列数据较少，因此是初步判断选择作用是否存在的首选检验方法。

计算 K_a 和 K_s 的方法有三类：以 Nei-Gojobori 为代表的进化通路法（Evolutionary Pathway Methods）(Nei & Gojobori, 1986)，以 Li-Wu-Luo 为代表的基于 Kimura 双参数模型的方法（Methods Based on Kimura's 2-Parameter Model）（Li et al, 1985），和以 Yang 的密码子替代模型为代表的最大似然法(Yang & Bielawski, 2000)。其中后两种方法比较常用。通过上述方法计算出 K_a 和 K_s 后，构建 Z 检验：

$$Z = \frac{K_a - K_s}{\sqrt{V(K_a - K_s)}} \quad (12)$$

（Nei & Kumar, 2000; 中译本见吕宝忠等[①]）

若（12）显示显著的结果，则暗示选择作用的存在。

3　结论与展望

中性检验以中性假说为统计学零假设，以检验 DNA 水平上选择作用是否存在为目的，在原理和统计检验的构建方面都具有严密的逻辑体系。但如前文所述，自然界中不同的进化因素将产生相同或相似的 DNA 多态结果。而且选择作用往往没有强烈到在基因序列上留下显著的印迹，或者由于其他的一些因素（如重组、后续中性突变等）而导致检验不显著。因此，在面临实际的数据时，现有的选择检验方法还有很多问题。检验所隐含的假设和有可能造成检验结果显著的选择以外的进化因素都会引起问题。例如许多检验方法都假设不改变氨基酸序列的同义突变为中性突变，但在某些序列区域，密码子偏倚（codon bias，在编码同一种氨基酸的数套密码子中，生物体倾向使用特定的某一套或两套密码子的现象）将使该假设不成立，因此会对检验结果产生影响。又如在某些检验（HKA 检验）中，要求所研究种群大小保持不变等等。

其次，单独某一个检验得到显著的结果有时并不足以表明选择的存在。实际操作时，通常先采用 K_a vs. K_s 检验比较序列同义突变和错义突变速率差异，然后通过 Tajima's D 检验进行种内多态性比较，最后用 MK 检验进行种间差异比较。通过几个检验方法才能对问题作出解释。此外，本文介绍的六个检验都是对中性假说模型的“保守”检验，原因是这些检验体现的只是自然选择对整个 DNA 序列作用的平均结果，而不能体现在不同区域和位点选择作用强弱的差异。假如在编码氨基酸的 DNA 序列的某一个位点发生了变化，该变异有可能改善了整个蛋白质的性质和功能，但中性检验却有可能无法分辨这样单个的序列突变，从而对检测这样的选择过程无能为力。

上述种种问题对进化遗传学家提出了挑战。当前，理论进化遗传学家的一个重要任务就是思考如何修正模型以符合 DNA 数据的实际情况。对此 FU Yun-xin 教授的实验室正在针对实际 DNA 数据的特性进行研究和探索（Fu, personal communication），并已取得了一些有意义的结果（Li et al, 2002）。

除了对有关理论模型进行修正，对于实际的数据还可以采取其他策略，如我们在雄性生殖基因中所作的那样通过电脑模拟比较观测到的变异格局是否显著偏离中性模型下的分布来检测选择的存在(Wyckoff et al, 2000)；或是判断氨基酸的改变是否显著改变了蛋白质的功能等等来推断选择作用是否存在（如 Zhang et al, 2002）。

致谢： 感谢 FU Yun-xin（符云新）教授的有益指点和讨论。感谢石宏博士和马普小组工作人员杨爽、李莎、张越和赵若萍在本文修改过程中给出的宝贵意见。

[①]分子进化与系统发育. 吕宝忠，钟　扬，高莉萍译. 北京：高等教育出版社. 47-48，218-231.

参考文献:

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection [J]. *Genetics*, **155**: 1405 - 1413.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila* [J]. *Nature*, **415**: 1024 - 1026.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection [J]. *Genetics*, **147**: 915 - 925.

Fu YX, Li WX. 1993. Statistical tests of neutrality of mutations [J]. *Genetics*, **133**: 693 - 709.

Hubby JL, Lewontin RC. 1966. A molecular approach to the study of genic heterozygosity in natural populations: Ⅰ. The number of alleles at different loci in *Drosophila* [J]. *Pseudoobscura Genetics*, **54**: 577 - 594.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination [J]. *Theor. Popul. Biol.*, **23** (2): 183 - 201.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data [J]. *Genetics*, **116**: 153 - 159.

Kimura M. 1968. Evolutionary rate at the molecular level [J]. *Nature*, **217**: 624 - 626.

Kimura M. 1983. The Neutral Theory of Molecular Evolution [M]. Cambridge: Cambridge University Press.

Kingman JFC. 2000. Origin of the coalescent: 1974 - 1982 [J]. *Genetics*, **156**: 1461 - 1463.

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster* [J]. *Nature*, **304** (5925): 412 - 417.

Kreitman M, Hudson RR. 1991. Inferriag the evolutiouary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergeuce [J]. *Genetics*, **127**: 565 - 582.

Li H, Zhang Y, Zhang YP, Fu YX. 2003. Neutrality tests using DNA polymorphism from multiple samples [J]. *Genetics*, **163**: 1147 - 1151.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes [J]. *Mol. Biol. Evol.*, **2** (2): 150 - 174.

Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila* [J]. *Science*, **260**: 91 - 95.

McDonald J, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila* [J]. *Nature*, **351** (6328): 652 - 654.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions [J]. *Mol. Biol. Evol.*, **3**: 418 - 426.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics [M]. Oxford: Oxford University Press. 258 - 264.

Ohta T, Kimura M. 1971. Behavior of neutral mutants influenced by associated overdominant loci in finite populations [J]. *Genetics*, **69** (2): 247 - 260.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data [J]. *Genetics*, **141**: 413 - 429.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations [J]. *Genetics*, **105**: 437 - 460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism [J]. *Genetics*, **123**: 185 - 195.

Wang W, Zhang JM, Alvarez C, Llopart A, Long M. 2000. The origin of the *jingwei* gene and the complex modular structure of its parental gene, *Yellow Emperor*, In *Drosophila melanogaster* [J]. *Mol. Biol. Evo.*, **17**: 1294 - 1301.

Wang W, Brunet FG, Nevo E, Long M. 2002a. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster* [J]. *PNAS*, **99**: 4448 - 4453.

Wang W, Thornton K, Berry A, Long M. 2002b. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome [J]. *Science*, **295**: 134 - 137.

Watterson G. 1975. On the number of segregation sites [J]. *Theor. Popul. Biol.*, **7**: 256 - 276.

Wyckoff G, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man [J]. *Nature*, **403**: 304 - 309.

Yang ZH, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation [J]. *Trends in Ecology and Evolution*, **15**: 496 - 503.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey [J]. *Nat. Genet.*, **30**: 411 - 415.